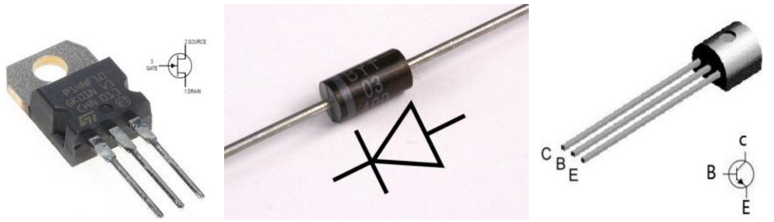
DIODES AND TRANSISTORS AS PHASE TRANSITION ELEMENTS

Gopi Vijaya



Introduction

Diodes and transistors form the basics of all technology today, and it is important to see how they really function, with the aid of the Reciprocal System. The basic constituents of these devices are semiconductors. It has already been described in an earlier paper [1] that semiconductors are substances containing elements with neutral valence and a capacity to convert the normally 1D electron motion to a 2D motion, as a phase transition (such as solid to liquid). The discussion in this paper is a continuation of that subject, and uses the terms doping, electrons, holes, and band-gaps in that sense. Electrons are 1D motions, while holes are 2D motions, and both are physically effective.

1. Standard Diode Theory

When a substance is "doped" as both p-type (hole excess) and n-type (electron excess) and joined together, it forms a *p-n junction*. This junction is at the heart of the device action of both diodes and transistors. However, historically the engineering was done well before a theory was ready to understand this process, and therefore the theory had to play catch up quite a lot[1]. As a result, there are considerable holes in the explanation provided, since the theory was put together with the concepts available at the time.

For instance, the standard diagram for the behavior of a p-n junction is this one [2]:

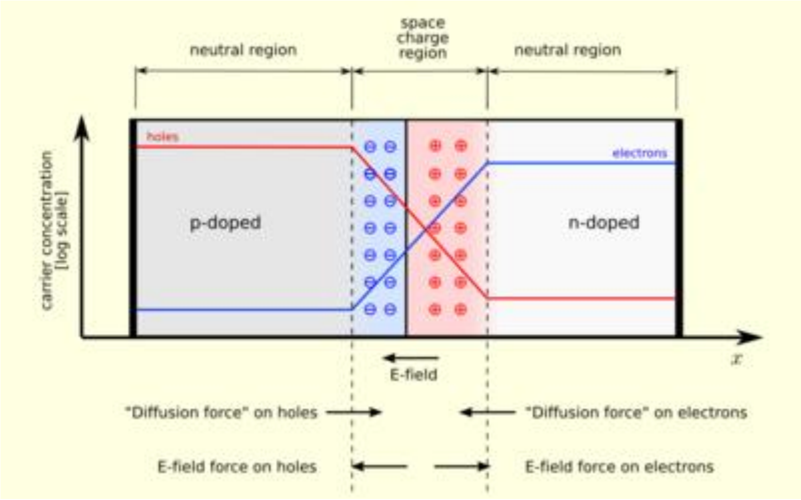

neutral region
space charge region
neutral region
holes
electrons
p-doped
n-doped
carrier concentration [log scale]
x
E-field
"Diffusion force" on holes
"Diffusion force" on electrons
E-field force on holes
E-field force on electrons

Fig.1: The P-N junction

[1] "We have seen how, in the prewar period, theorists advanced from one unreliable hypothesis to another like travelers in a swamp who step from one tussock to another, abandoning each just as it sinks, yet not without making some forward progress" – L. Hoddeson, *Out of the Crystal Maze*, Oxford University Press, New York (1992), p. 300

In this picture, the positive (hole excess) material is on the left and the negative (n-excess) material is on the right. The junction is formed when the two are contiguous, and at the junction itself, there is equilibrium. It is said that electrons and holes "diffuse" in opposite directions, and end up creating an electric field that opposes further flow of them. This explanation, however, raises several questions: if there is a tendency of creating an electrostatic field, which is one of the most powerful fields that exist, why would the electrons or holes ever diffuse? If they do start diffusing, why would they ever stop until the whole slab is neutralized? Besides, if holes "do not really exist," as is often claimed, how would an electron alone create a field with itself? Since only a force can balance a force, what is the force that is called "diffusion force" in the diagram? Is it something that is different from electrostatic? How can concentration of holes or electrons create a force? How can they both move around as if the nuclei of the atomic lattice did not exist at all? The more one looks at the diagram, it becomes clear that the sort of particles imagined to exist can never physically exist, without introducing a large number of arbitrary assumptions. It is therefore very important to understand the nature of the p-n junction clearly, which will enable a clear understanding of diodes and transistors as well.

2. Conceptual background

The primary feature to understand is that there is a difference in the structure of holes and electrons: holes are 2D while electrons, at least to start with, are 1D. Therefore, when a "p-region" is sandwiched to an "n-region", the two states or phases are physically next to each other. Hence a 2D magnetic motion-rich region is adjacent to a 1D electric motion-rich region. The two "carriers", i.e. *uncharged* electron and holes, behave differently in this system. Electrons are 1D, and they can get transferred to 2D if there is availability of the bandgap energy. Holes are 2D to begin with, and they can get transferred to 1D only with difficulty. Therefore, there is a possibility of electrons moving from 1D to 2D, with the addition of a bandgap boost. There is a similar (much smaller) possibility of holes moving from 2D to 1D. But if there is ambient thermal energy available to do the job, this transition does happen, and creates a thin junction. This is the "diffusion". *The magnetic dimension is less sensitive to the electric potential difference, so the diffusion is not stopped in spite of there being a potential against it.* On one side of the junction, 1D electrons (n region) receive the 1D holes, while on the other side, 2D holes receive the newly converted 2D electrons. However, when the transition is actually done, the electron enters a domain where the hole normally is, and the hole enters a domain where the electron normally is. This results in neutralization of motion in both cases, which prevents further movement. This is the "opposing electric field" which creates a thin region which is called the "space charge region" in the diagram.

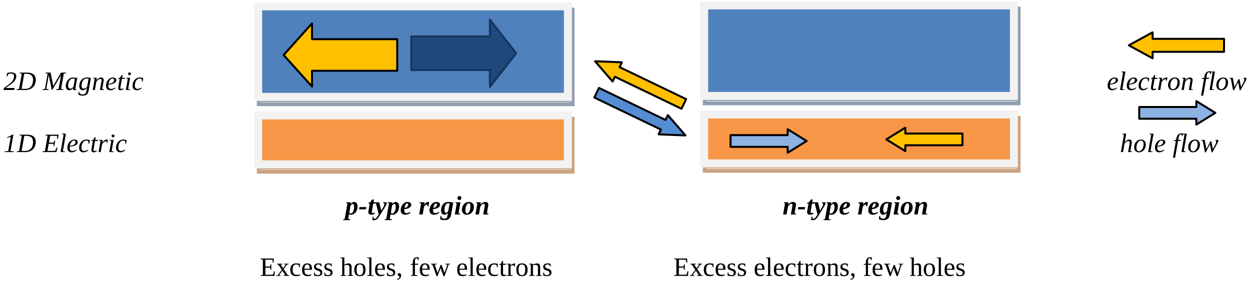


Fig. 2: The p-n junction equilibrium

Hence, the so-called diffusion and field responses are actually due to the interaction of *two parallel processes*. This removes the confusion of these two processes in conventional understanding, and provides answers to a whole host of problems. For instance, it is known that hole-mobility is in general much smaller than electron-mobility, and no clear reason is provided why this should be so. However, since the hole is 2D, and electron is 1D, the hole can be resisted much more effectively than the electron. The equilibrium between the n-side and p-side is not due to the same electron getting affected by two different forces at once, but it is due to the presence of an electric and a magnetic pathway, with different dynamics for each. That also explains why the hole or electron "diffuse": their movement is based on the 2D to 1D or 1D to 2D transition which occurs as long as the transition energy is available, while the “field” acts only in the 1D zone or 2D zone alone. If the entire process was only to be carried out with the qualitatively identical electrons and holes as in conventional theory, the process would stop even before it began.

What happens when an electric potential is applied to the system, in what is usually called “forward bias”? There is an increase in the pressure on the electrons to move, and the excess electrons available transition from 1D to 2D. There is hence an easy flow of electricity. The hole current arises only as a reaction to all the electrons entering the 2D mode, and the only place it can move into is the excess-electron 1D displacement of the n-region, since time/space is motion. Therefore the electron current is from n-to-p while the resulting hole current, in the 2D channel, is from p-to-n. This is exactly what is required for the normal forward operation of a diode.

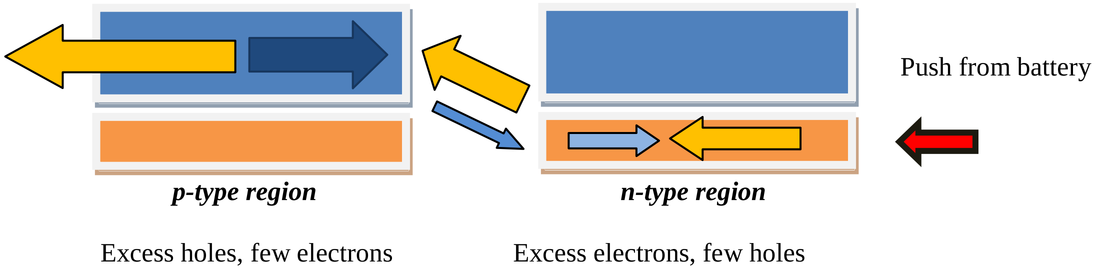


Fig. 3: The p-n junction in Forward Bias

And what happens when a reverse potential is applied as “reverse bias”? Since the push of the battery is *mainly in the electric dimension*, there will be an attempt to push electrons from the p-side to the n-side. However, there is a severe lack of electrons on the p-side both on the 1D electric and the 2D magnetic dimensions, and in addition, whatever electrons are available have to “squeeze” from 2D to 1D in order to transition into the n-side. This creates a block, and since the reduction of an electron current also means the reduction of a hole current, there is barely a trickle of current in the system. Thus the diode stops current. The dimensional asymmetry is hence crucial to the process of rectification.

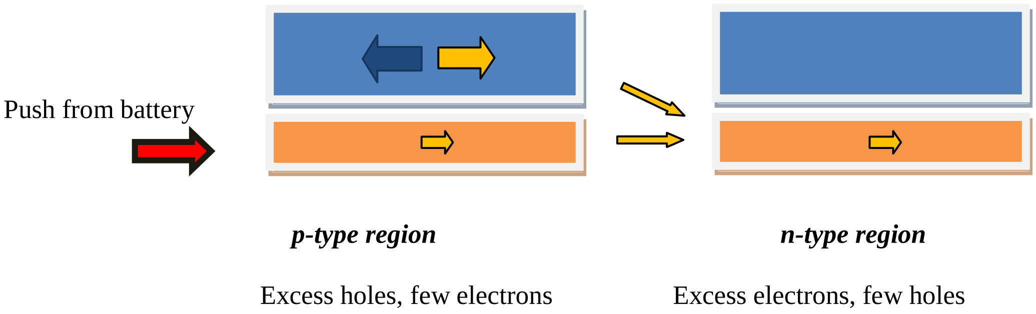


Fig. 4: The p-n junction in Reverse Bias

3. A Can of Boiling Water

There is a very useful analogy that can be made, by taking help of a different domain in the Reciprocal System. In the liquid state, there is a speed that removes gravitational motion in one-dimension, while in the vapor state, the gravitational effect is removed in two dimensions [3]. Therefore a liquid-vapor system offers a perfect analogy to a 1D-2D transition system, which clarifies the concepts with clear visuals.

Imagine a container containing water or any other liquid, sealed all around except for a small inlet in the bottom for water and one outlet at the top. Now imagine that the air in the container is saturated with water vapor, because the entire system is present in a hot environment.

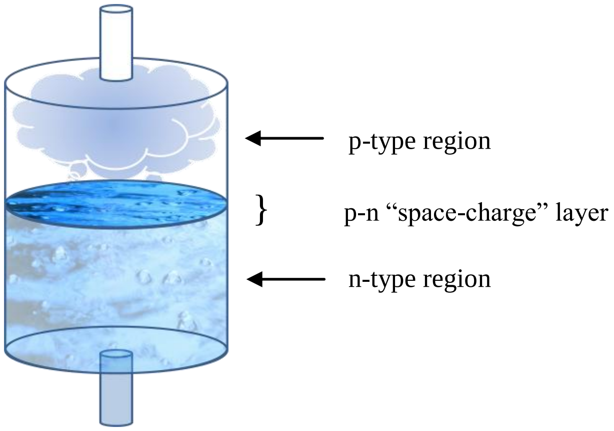


Fig. 5: The can of boiling liquid

Since the liquid has 1D of freedom, and the vapor has 2D, the liquid region is the “n-type” and the saturated vapor region is the “p-type” region. The liquid is electron in 1D form, while the vapor that forms is electron in 2D form, with the holes being the gaps in vapor. In case of equilibrium, there are two events happening at the liquid-vapor interface. Liquid is evaporating, and vapor is condensing. Some electrons are evaporating into the hole region as droplets, and some holes are condensing into the electron region, as bubbles. Now consider what happens when this system is “forward biased” by extra heating *from the bottom*. Heat is the equivalent of voltage. The liquid boils off some more, and presses through the tube at the top, which creates an *electron current*. Now, suppose that the system is “reverse biased”, by heating *from the top*. This will increase the pressure of the water vapor, but it does not lead to any current because it is far more difficult for vapor to shoot into the liquid surface downward and push through the lower outlet. Hence, this system conducts “electricity” only in *one direction* – namely, as a diode.

The key requirement for this process is the latent heat of evaporation, which is the amount of heat needed to convert liquid into vapor. When there is a phase transition involved, an exponential dependence of the vapor pressure on the temperature is expected. Sure enough, this is what is obtained in determinations of vapor pressure [4]:

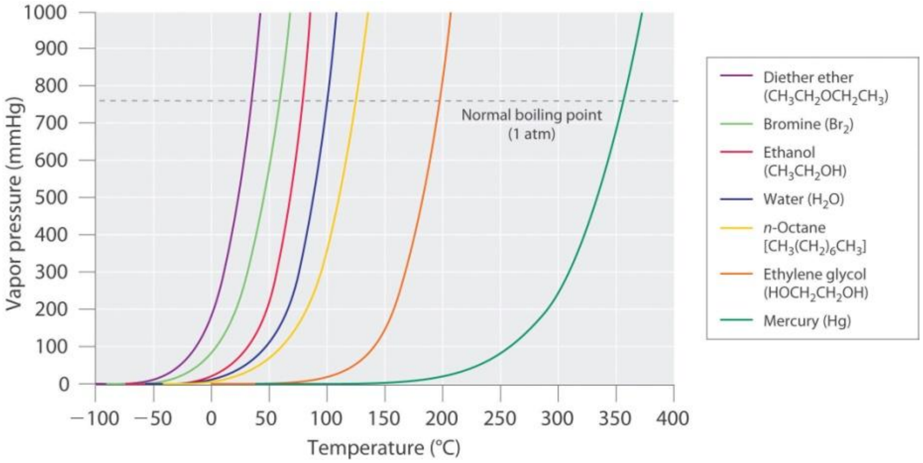


Fig. 6: Vapor-pressure vs Temperature for a number of liquids

This is very similar to the diode characteristics shown in figure 7 below.

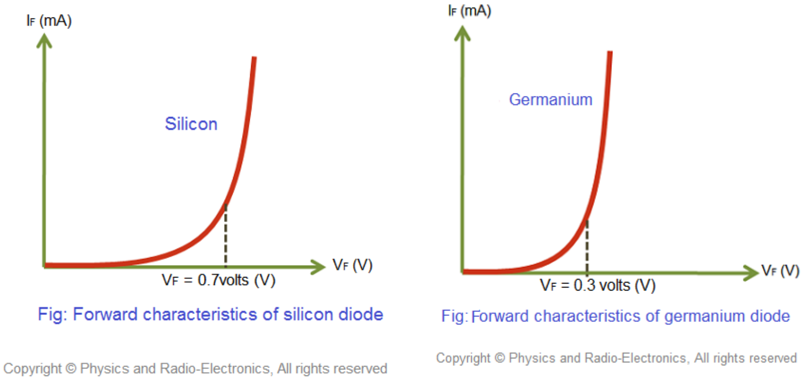


Fig: Forward characteristics of silicon diode

Copyright © Physics and Radio-Electronics, All rights reserved

Fig: Forward characteristics of germanium diode

Copyright © Physics and Radio-Electronics, All rights reserved

Fig. 7: Current-Voltage characteristics of silicon and germanium diodes

Thus, the behavior of a diode clearly shows the signature of a phase transition, matching with the well-known phase transition of liquid to vapor. And when the boiling liquid "diode" is reverse biased, then the vapor pressure is increased until the container bursts due to the pressure, which is *exactly* what happens in a junction breakdown: the diode fails.

Before extending this analogy to the next device, the transistor, there is an additional feature that has to be clarified.

4. The Product Rule

In the prior description of the p-n junction with the arrows, the thickness of the arrows was used as a surrogate for the magnitude of motion. However, there is an exact rule that governs the number of electrons and holes in each substance. This has been briefly mentioned in the earlier paper [1], but a fuller treatment is given here.

The origin of electrons and holes is not the material itself, but the surroundings of the material. It has already been mentioned that there is generally a flux both of uncharged electrons 0-0-(1) and holes ½-½-0 in the environment, and since they do not have sufficient rotation in all three dimensions, they proceed with the progression of the natural reference system at the speed of light. However, when there are specific biasing conditions, like a battery, a part of this flux gets channeled into the electric displacement or magnetic displacement. Since the conditions within the material determine whether the electrons or the holes are routed through, it is a property of the material that decides the numbers of electrons and holes. Which means the way space and time are related determines these numbers.

Motion is conserved in any system, therefore the conservation rules apply for all displacements in the system. In case of the regular undoped material in the non-conducting state, there is no outside influence for determining the number of current carriers. Therefore, in order to retain equilibrium, the space displacement has to equal the time displacement. Since time displacement controls the number of electrons and space displacement controls the number of holes that can pass through, we have:

$$s = t \tag{1}$$

$$p = n \tag{2}$$

This holds true even when heat is added to the system, since the heat is a vibration it increases the numbers of *both* electrons and holes, maintaining the equality. However, when there is an external disturbance of any sort, such as dopants, pressure, or a chemically reactive gas, this equilibrium gets disturbed. Either the space displacement or the time displacement changes value. The magnitude of space displacement cannot however exceed the amount of equivalent space that is provided by the temporal structure of the substance, and the relationship of equivalent space to time is:

$$s = \frac{1}{t} \qquad (3)$$

Therefore, translating this into electron and hole magnitudes:

$$p = \frac{1}{n} \qquad (4)$$

$$np = 1 \qquad (5)$$

The PRODUCT of electrons and holes in any substance is hence a constant, and this is derived from the fundamental properties of space and time when forming matter. This principle, which can be called the product rule, is a different relation from equation (2), and hence has a different behavior. Conventional science has a long derivation of this principle, called the "mass action law", going through a whole series of steps which make use of the empirically determined exponential relation [5]. However, in the Reciprocal System, it is a straightforward property of all material systems. *The exponential law is simply a different expression of equation 5, since products become sums when one raises them to exponents.*

It is also counterintuitive for the traditional "object" view of the world to account for this sort of change. The usual explanation for a hole is that it is the absence of an electron. A diagram like this is usually given to show how a hole moves:

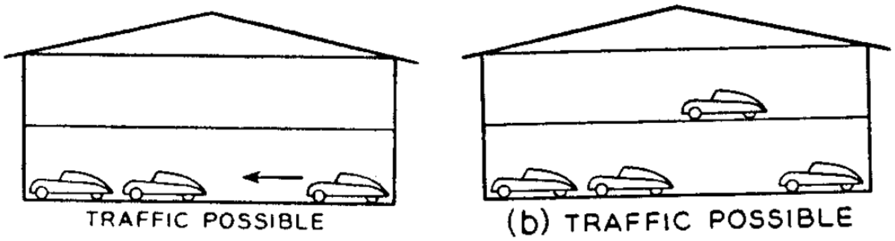


Fig. 8: As the cars (electrons) move forward, the absence of a car (hole) moves backward. (b) If an electron is moved to a different level, a hole is created. From [6].

For instance, it is easy to see that the number of electrons must equal the number of holes due to symmetry. If one car moves, one hole moves. If one car is moved up, a hole is created in the lower level. However, one cannot say why, when a bunch of electrons are added to the system, the system "loses" precisely the number of holes needed to keep the PRODUCT a constant. If there are a few new cars added to the lower level, for instance, how do a precise number of cars move to a higher level in order to keep the product of "cars" and "absent cars" a constant? How can there even BE a product of "cars" and "absent cars"? And on the other hand, if more holes are created by moving cars to the upper level, how do a certain number of cars disappear entirely, in order to correspondingly decrease the number of cars? In spite of basic logical problems like this, following the pioneering work of Shockley [6], all electronics textbooks and research carry some version of this description in their explanations.

In the context of the Reciprocal System, both electrons and holes are in plentiful supply, even if holes are a little less plentiful than electrons in a material environment, and more importantly, *they are motions*. Hence when the number of electrons or holes is altered in a substance, the required number of electrons and holes are drawn in by the environment or expelled into it. In addition, there is a clear explanation for two levels shown in Fig. 8: the lower level is the electric displacement, and the upper level is the magnetic displacement, the two aspects of a material system. *This transfer between these levels is the basis for all "band diagrams."*

The analogy of the vapor pressure fails to capture this property, as it does not have the product of the amount of vapor and the amount of liquid a constant. This feature is specific to atomic-level systems, and therefore has to be dealt with at that level. The product rule, or equivalently the exponential relation is, however, extremely important to understand the most ubiquitous device in current electronics: the transistor.

5. Transistor Action

The traditional bi-junction transistor (BJT) is a device that makes use of a back-to-back p-n junction, either in the form of n-p-n or p-n-p. In this system, one p-n junction acts as a controller, like a tap opener. The other junction allows the flow of current. The working of the transistor is attributed to a process called "minority carrier injection" which will be described shortly. Majority carrier is the name for electrons (1D) in an n-type and holes (2D) in a p-type material. Minority carrier is the name for holes (1D) in n-type material and electron (2D) in p-type material.

The general structure of a commonly used transistor is the n-p-n version:

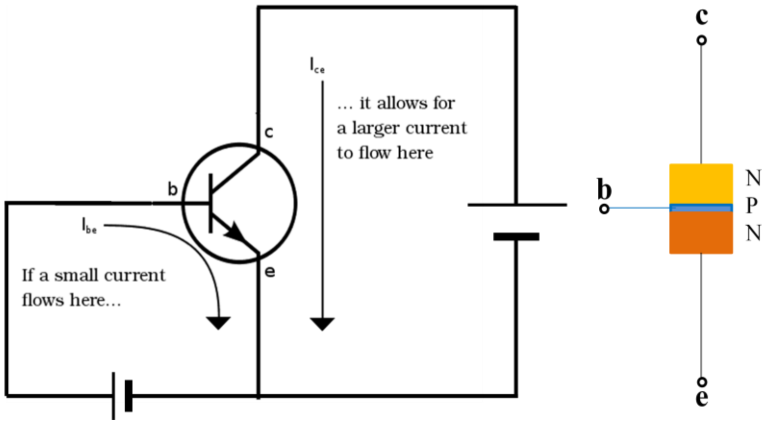


Fig. 9: From [7], the circuit on the left controls the current in the circuit on the right. The n-p-n structure is shown in color, and the lower darker colored n-type (n-region-I) is more heavily doped than the upper one (n-region-II). Current direction is for holes, electron current is opposite.

The connections "e", "b" and "c" will be used purely as labels in this discussion. The action of the transistor consists of the following steps:

1. A small voltage is applied on the left circuit between the p-n junction, in forward bias mode b-e.

2. A large voltage is applied on the right circuit between "c" and "e".

3. This leads to a current in the left circuit, with electron flow anticlockwise from n to p, from "e" to "b".

4. The electrons that end up in the thin p-junction have two routes: the left and the right circuit. Some go to the left into "b," some go straight onto the right circuit along "e"-"c".

5. The current that goes straight through into the right circuit is much larger than the one on the left. This leads to a current amplification in the circuit with the higher voltage on the right.

Steps 1-3 are clear enough to understand, however it is in step 4 and 5 that questions arise. How does a current flow arise at all in the right hand circuit when one of the junctions is reverse biased and one is forward biased? More importantly, how is the current enhanced, instead of being diminished by the reverse bias? How does the electron current couple with the second junction? Why does one need an n-type material for the "c" instead of just connecting up a conductor? The conventional explanation says that electrons "diffuse" into the top junction "c" from the bottom junction "e" via "b". How can they diffuse, *into* a reverse biased system, like water flowing uphill? How come the reverse voltage at "c" does not stop the diffusion? Does the voltage not affect the electrons somehow? If so, why not? Why is there a reason for the right hand circuit to even be activated? And when the whole thing is activated, why does a very small current pass through the left circuit, and a large one through the circuit on the right?

The conventional explanation fails because it attributes two different behaviors simultaneously to the same current once again: one of *drift* and one of *diffusion*. The crucial difference between the two is that drift is mainly active in the electric displacement channel, while diffusion is active in the magnetic displacement channel. It is hence extremely important to keep the concept brought out in Section 1 in this context: *The magnetic dimension is less sensitive to the electric potential difference, so the diffusion is not stopped in spite of there being a potential against it.* An additional fact is that, 2D←→1D transition involves exponentials, as already described in Section 2. So it is to be expected that in a p-n junction, there is an exponential decay of the concentration of carriers (electrons in this example) of this form, with or without bias:

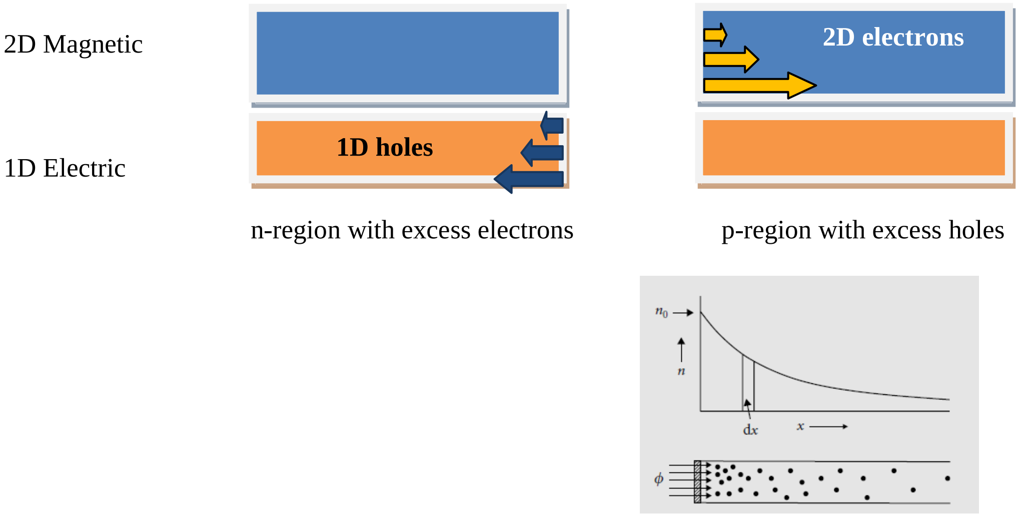


Fig. 10: Same condition as Fig 2, with p and n exchanged, showing the concentration gradients of 2D electrons and 1D holes (both "minority carriers") on both sides. The relation between 2D and 1D gives the exponential gradients

Now consider what happens when the p region is much thinner:

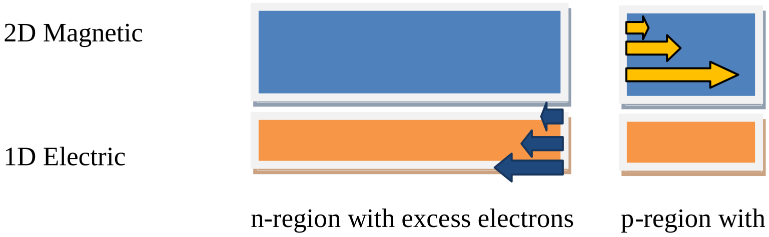


Fig.11: Thin p-region allows a greater chance of electronic concentration penetrating to the rightmost surface.

In this case, there is an inherent chance for the electrons in the p-region to be closer to the rightmost edge. Now, when there is another n-junction (II) to the right with a lighter doping than the one on the left (I), the entire set up is of the form:

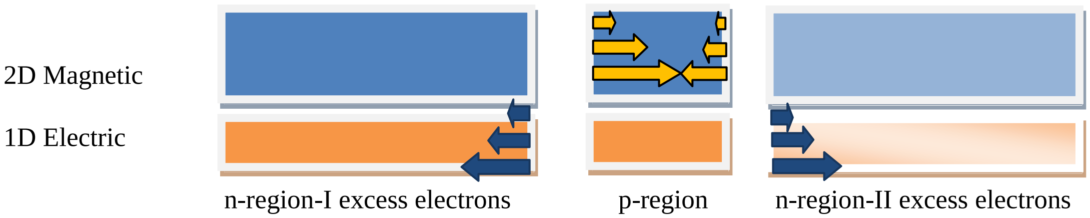


Fig. 12: An n-p-n region in equilibrium, showing how the minority carrier in each region has its concentration distribution. In the n-region, it is the 1D hole. In the p region, it is the 2D electrons.

If the p-n junction on the left is forward biased, and the one on the left is reverse biased, what is expected to happen?

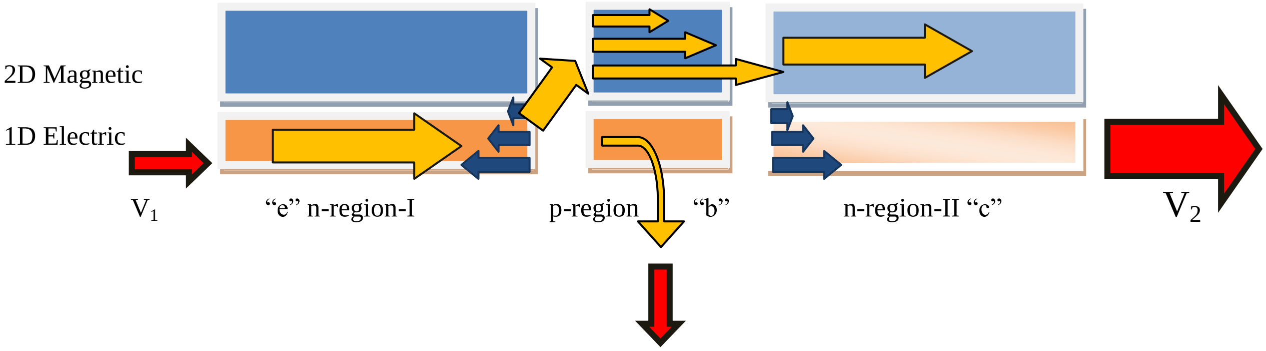


Fig. 13: An n-p-n region with a forward bias applied in the junction on the left (across "e-b") and a negative bias applied to the system from the right (between "c" and "e"), just before the system responds

There are several things to keep in mind. Recall from section 2 that the push from the battery acts *mainly in the electric dimension*. A forward bias V_1 in the junction on the left shifts the majority of excess electrons in the n-region-I into the 2D magnetic displacement of the p-region. The situation is similar to figure 3. In the reverse biased junction on the right, the situation is markedly different from that shown in figure 4. Unlike the regular reverse-biased junction that has very few 2D or 1D electrons in the p-region, this time the p-region does contain 2D electrons swept in from the forward bias V_2. Keeping in mind the requirement that the push of a battery is less effective in the magnetic dimension, one needs a larger voltage *and* a greater doping of the n-region-I for the electron supply. Thus, the 2D electrons get swept to the right by the battery. The amount of 2D electrons available in the p-region increases exponentially with the voltage. Therefore, changing the smaller voltage in the "e-b" circuit controls the magnitude of the resulting current very effectively. Lastly, a small percentage of the electrons swept in from the left do enter the electric displacement of the p-region, where the voltage V_1 is effective through the conductor at "b". They cannot go through the electric displacement of the n-region-II, since that has an excess of electrons and space/space is not motion. The only available path for these electrons is out through "b". So only a few electrons complete the circuit on the left, making this current very small. The voltage in the "e-b" circuit hence generates a large current in "c-b" and a small current in "e-b". This completes the transistor action.

With this functioning in mind, it easy to see why the names given to "e", "b" and "c" are emitter, base and collector. The highly doped emitter emits uncharged electrons, the collector collects this current, while the base modulates it. The transistor therefore contains both a 1D and a 2D electron current (the hole current is complementary to electron motion). The 1D circuit is the emitter-base circuit, while the 2D circuit is the emitter-collector circuit, or more accurately, the base-collector part of the emitter-collector circuit. While normally, a reverse biased junction does not allow any current, sandwiching a forward biased p-n junction behind the reverse-biased one populates the 2D magnetic displacement, allowing a current through the system. It can be seen that a strong doping of the p-region results in a good supply of excess holes ½-½-0, which allows a better flow of 2D electrons (½)-(½)-0 through them. Even a metal will suffice for this action, and it is seen that n-metal-n regions also generate transistor action [8]. The usually unexplained difference of magnitude between the small current in emitter-base circuit and the large current in the emitter-collector circuit is now seen to be due to the different populations of electrons in the 1D and 2D displacements of the p-region. The lightly doped n-region-II is also extremely important, not only because it helps generate the reverse bias that keeps the system from having a current when the p-n junction is not yet forward biased (i.e. when it is in the off state), but it also helps to isolate the two batteries from each other. Without it, the batteries would fight each other and short out.

The key to transistor action is hence the transfer of resistance across from the 1D to the 2D region: giving a clear meaning to its name as well (**transist**or = **trans**fer of res**ist**ance). Discussion of the field effect transistor (FET), which accomplishes the same process using fields perpendicular to the emitter-collector current instead of the base current, will be deferred to another paper.

Summary

It was shown that the standard theory for both diodes and transistors attempts to use only one sort of current – the 1D electron charge – to account for two different behaviors seen in the system: *drift* and *diffusion*. It has been identified that there are two currents for electrons and two currents for holes, (1D and 2D) which determine two parallel systems of currents. One of them is in the electric displacement and is more sensitive to external voltage (equivalent to drift), while the other is in the magnetic displacement and is less sensitive (equivalent to diffusion). The two currents for electrons, for instance, have entirely different dynamics, and it is important to understand both of them to clarify the behavior of a p-n junction or diode.

The transition from 1D to 2D for electrons can be understood using the analogy of the boiling liquid under vapor pressure. It indicates the presence of exponential relations, as well as the product rule, wherever the 2D←→1D transitions occur. This analogy is a good one to understand the behavior of a diode, up to a point. The two different channels available for carrier motion also make it easier to understand the functioning of a transistor, where the "majority carriers" and "minority carriers" for electrons are simply their 1D and 2D versions respectively. This clarification resolves contradictions that abound if the entire explanation is based only on a single charge carrier.

Bibliography

[1] G. Vijaya, "Semiconductors and Doping in the Reciprocal System," Salt Lake City, 2018.

[2] Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/P%E2%80%93n_junction.

[3] "States of Matter," [Online]. Available: https://library.rstheory.org/books/nlosat/09.html.

[4] [Online]. Available: https://catalog.flatworldknowledge.com/bookhub/reader/26669?e=averill_1.0-ch11_s04.

[5] [Online]. Available: http://www.tf.uni-kiel.de/matwis/amat/semi_en/kap_2/backbone/r2_2_2.html.

[6] W. Shockley, "Donors and Acceptors," in *Electrons and Holes in Semiconductors*, Princeton, NJ, D. Van Nostrand Company, 1959, p. 14.

[7] [Online]. Available: https://www.build-electronic-circuits.com/how-transistors-work/.

[8] "PPPPPPPPPPPS, point 4.," [Online]. Available: http://amasci.com/amateur/transis2.html.